IN THE CLAIMS

Please amend the claims as follows:

1. (CURRENTLY AMENDED) A server comprising:

a processor;

a memory;

a system area network connection;

a local area network connection;

wherein the processor, memory, system area network connection, and local area network connection are operably interconnected within the system; and

software held in the memory and operable on the processor to:

load unique content into the memory from a storage location,

receive requests for content over the local area network connection,

service requests for the content in memory,

service requests for content located in a memory of another server by obtaining the content over the system area network connection, wherein the another server is identified as a function of a table holding content availability and location data of content held in memory of one or more other servers; and

cache content used to service request for content located in the memory of the other server for use in servicing subsequent requests for identical content.

2. (Original) The server of claim 1, wherein the software operable on the processor is a component of an operating system of the server.

3. (Original) The server of claim 1, wherein the software operable on the processor is a driver.

4. (Original) The server of claim 1, wherein the software operable on the processor is a middleware component.

5. (Original) The server of claim 1, wherein the system area network is a Gigabit Ethernet network.

6. (Original) The server of claim 5, wherein the Gigabit Ethernet network is TCP Offload Engine enabled.

7. (Original) The server of claim 1, wherein the unique content is loaded into memory prior to the server being available to service content requests.

8-14. (CANCELLED)

15. (Original) A method of server operation comprising:

priming a memory of a server, wherein the server is a member of a server cluster, wherein the content in the memory of the server is unique to the server amongst all servers in the server cluster;

making the content in the server memory available to other servers in the server cluster over a high-speed interconnection;

receiving requests for content;

fulfilling content requests by retrieving data from the server memory and from memories of one or more other servers over the high-speed interconnection; and

caching content of other servers that has been requested either recently or commonly to provide the server the ability to fulfill requests for cached content locally.

16. (Original) The method of claim 15, wherein requests for content are received over a local area network connection.

17. (Original) The method of claim 16, wherein requests are received into the local area network on a router coupled to the Internet.

18. (CURRENTLY AMENDED) A method comprising:

distributing web content across a cluster of web servers connected by a first network;

fetching, by a first one of the web servers, web content on demand from a memory of a second one of the web servers in the cluster of web servers across the first network; and

caching the web content in the memory of the first one of the web servers.

19. (CANCELLED)

20. (Original) The method of claim 18 further comprising responding to the request with the web content from the memory of the first server.

21. (CURRENTLY AMENDED) An article comprising a physical computer-readable medium containing associated information, wherein the information, when accessed, results in a machine performing:

receiving, by a first server in a plurality of interconnected servers, a request for content; and

determining if the content is available in a memory of the first server:

if the content is available in the memory of the first server, then responding to the request with the content from the memory of the first server; and

if the content is not available in the memory of the first server: ~~, then~~

identifying another server of the plurality of interconnected servers holding the requested content in memory as a function of a table holding content availability and location data of content held in memory of each of the plurality of interconnected servers; and

obtaining the content from [[a]] the memory of ~~one of~~ the ~~servers in~~ another server of the plurality of interconnected servers ~~other than the first server~~ and replicating the content in the memory of the first server.

22. (Original) The article of claim 21 further comprising responding to the request with the content from the memory of the first server.

23. (Original) The article of claim 22 further comprising responding to a subsequent request for the content with the content from the memory of the first server.